An Approach to Proper Name Tagging for German

Christine Thielen

Seminar für Sprachwissenschaft

Universität Tübingen

Wilhelmstr. 113

D–72074 Tübingen

Germany

e-mail: thielen@sfs.nphil.uni-tuebingen.de

Abstract

This paper presents an incremental method for the tagging of proper names in German newspaper texts. The tagging is performed by the analysis of the syntactic and textual contexts of proper names together with a morphological analysis. The proper names selected by this process supply new contexts which can be used for finding new proper names, and so on. This procedure was applied to a small German corpus (50,000 words) and correctly disambiguated 65% of the capitalized words, which should improve when it is applied to a very large corpus.

1 Introduction

The recognition of proper names constitutes one of the major problems for the wealth of tagging systems developed in the last few years. Most of these systems are statistically based and make use of statistical properties which are acquired from a large manually tagged training corpus. The formation of new proper names, especially personal names, is very productive, and it is not feasible to list them in a static lexicon. As Church (Church, 1988) already discussed for English, it is difficult to decide whether a capitalized word is a proper name if it has a low frequency (< 20), and so they were removed from the lexicon. But because they are highly individual, this is the case for most proper names. Furthermore, the problem of proper name tagging for German is not restricted to the disambiguation of sentence–initial words, because proper names and generic terms (normal nouns) are capitalized both at the beginning and within a sentence. Church

suggested labelling words as proper nouns if they are "adjacent to" other capitalized words. This also holds for German proper nouns, but it is difficult to decide which of the capitalized words belong to the proper name and which not, e.g. is it a first name (as in "Helmut Kohl") or is it an apposition (as in "Bundeskanzler Kohl"), or is it a complex institutional name composed of several generic terms and a proper name (as in "Vereinigte Staaten von Amerika"). In this procedure, I use Church's heuristic for the selection of proper name hypotheses, which are evaluated on the basis of their syntactic and textual context together with a morphological analysis. The starting point of the analysis is a small database of definite minimal contexts like titles (e.g. "Prof.", "Dr.") and forms of address (e.g. "Herr", "Frau"), which increases with the processing of texts in which proper names are identified, and supplies new contexts which can be used to find new proper names and new contexts, etc.. This incremental method is applied to unrestricted texts of a small corpus (50,000 words) of German newspapers.

2 Proper Name Acquisition

From a psycholinguistic point of view it is possible that we memorize proper names better if we organize them in a hierarchy, in which each word would constitute a node whose subordinate nodes are its hyponyms (Koss, 1990). For example, we find in the semantic hierarchy in figure 1 SOCRATES as hyponym of PHILOSOPHER and PHILOSOPHER as hyponym of SCHOLAR, and each node may bear features describing properties of the node.

One can observe that hyperonyms of names are used to identify or to introduce a proper name in texts. If the knowledge of a name cannot be
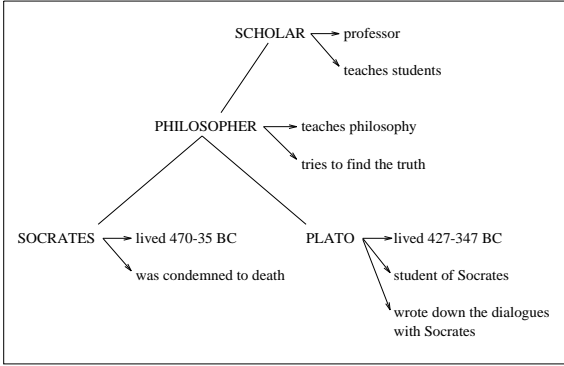
Figure 1: SOCRATES in a semantic hierarchy

presupposed, then the name is often introduced by an appositional construction (1)-(2) (Hackel, 1986) and can be used without additional information (3)–(4) (Kalverkämper, 1978) later on.

(1) der Vorsitzende des Verteidigungsausschusses, *Biehle* (CSU), hat Verteidigungsminister *Wörner* gebeten, ...
(the chair of the defence committee, *Biehle* (CSU), asked the Minister of Defence *Wörner* to ...)

(2) der SPD–Abgeordnete *Gerster* kritisierte, daß ...
(the SPD member of parliament *Gerster* criticized that ...)

(3) In einem Fernschreiben an *Wörner*, äußerte *Biehle* am Dienstag, ...
(in a telex to *Wörner*, *Biehle* commented on Tuesday ...)

(4) *Gerster* forderte eine Mindestflughöhe von 300 Metern
(*Gerster* called for a minimal flying height of 300 metres)

The syntactic analysis (see section 3) operates on a small lexicon of definite minimal contexts of proper names (MC–lexicon) which are used in such appositional constructions and generates a lexicon of so–called potential minimal contexts (MCpot–lexicon).

In addition there exist other methods (Koss, 1987) for the acquisiton of proper names, two of which can be directly observed in the texts. The first method ("Lernpsychologische Sinnverleihung") tries to lend sense to the name in order to learn it, e.g. the name "Düsseldorf" is given the meaning of 'village'. Today it is a big city, but the compound part *-dorf* helps us to identify it as a proper name. The second method, the formation of name fields ("Namenfelder") and

name scenes ("Namenlandschaften"), helps us to recognize names describing places which belong to a certain district or scenery, e.g., cities in the Stuttgart area like "Tübingen", "Reutlingen", "Esslingen" have the common suffix *-ingen*.

The morphological analysis (see section 3) operates with a list of so–called onomastic suffixes to identify place names.

3 Proper Name Tagging

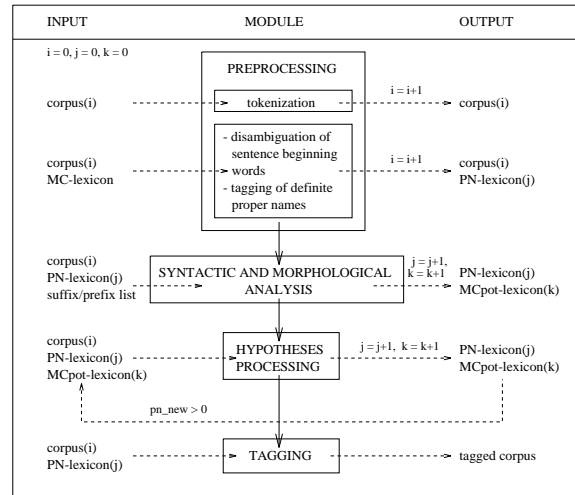An overview of the tagging process is shown in figure 2.

Figure 2: proper name tagging

Preprocessing

The corpus has to be preprocessed first of all. This includes the tokenization of the corpus in which all punctuation marks are separated from the words to allow the following disambiguation of sentence–initial words. This disambiguation uses a heuristic derived from the one used in CLAWS (Garside et al., 1987): if a sentence–initial word also occurs inside of a sentence with a lower case initial letter, then it is not a noun (normal noun or proper name) and represented with lower case letters. For this I use a list of all words with lower case initial letter found in the corpus which is stored in an AVL–tree (Wirth, 1983) for better searching and inserting.

After this, a first run through the corpus is done to identify definite proper names occuring in the contexts of the MC–lexicon. Apart from appositons as mentioned above, this lexicon contains speech–embedding ("redeeinbettende") verbs like "sagte"and "fragte" frequently used in political newspaper texts, as in:

(5) die Abgeordnete *Kelly* sagte, ...
(the member of parliament *Kelly* said, ...)

(6) *Heinlein* fügte hinzu, ...
(*Heinlein* added, ...)

(7) so fragte *Apel*
(*Apel* asked)

The MC–lexicon also contains prepositions and preposition frames to identify place names, as in:

(8) bei *Frankfurt*
(near *Frankfurt*)

(9) aus *Söllingen* bei *Baden–Baden*
(from *Söllingen* near *Baden–Baden*)

(10) im Raum *Landshut*
(in the *Landshut* area)

All proper names are stored in the PN–lexicon which is used during the entire processing.

Syntactic and Morphological Analysis

In the following analysis, the immediate syntactic and morphological context of all capitalized words is examined. If the capitalized word is already included in the PN–lexicon, then its immediately preceding context is stored as a potential minimal context in the MCpot–lexicon if it comprises one or more capitalized words. Cases where the proper name is marked as genitive are not considered because this could lead to wrong MCs (e.g., *Aussage Wörners, Besuch Lafontaines*). The collection of potential minimal contexts is also done in the hypotheses processing, which follows. For example, the proper name *Wörner* supplies the MCs: *Bundesverteidigungsminister, Verteidigungsminister, Minister, Nato–Generalsekretär*.

For the recognition of place names, a suffix list is used containing onomastic suffixes like *–acker, –aich, –beuren, –hafen, –hausen, –stetten, –weiler* and a prefix list containing prefixes like *Mittel–, Ost–, West–, Zentral–*. In addition to this the ending of the left capitalized word of two adjacents is checked for adjectival endings *–er, –aner*, as in:

(11) Mainzer Landtag
(the state parliament of *Mainz*)

(12) Münsteraner Parteitag
(the party conference of *Münster*)

If they also occur without this ending (*Mainz, Münster*), then these forms are proper nouns and

Node	List	Article
ADN	bei	0
	Nachrichtenagentur	
Angaben	nach	1
	Donnerstag	
Belgien	aus	0
	in	
Baum	FDP-Politker	0
	FDP-Abgeordnete	

Table 1: contexts of capitalized words

are stored in the PN–lexicon. The adjectival forms in (11)–(12) are considered as adjectives (following (Fleischer, 1989), p. 265).

Furthermore, loose appositional constructions ("lockere appositionelle Konstruktionen", (Hackel, 1986)) as in (13)–(14) are analyzed according to the patterns of noun phrases which occur before the proper name.

(13) der Staatssekretär des Landesinnenministeriums, *Basten*, ...
(the under–secretary of the Department of the Interior, *Basten*, ...)

(14) der Chef des Schweizer Wehrministeriums, Bundesrat *Koller*, ...
(the director of the Swiss Department of the Armed Forces, the minister of state *Koller*, ...)

During this run through the corpus, a second AVL–tree is constructed in which all capitalized words are stored together with some information that can be useful for the hypotheses processing. For each word (node) there is a counter for all occurences of the word with an article and a list of all its immediately preceding words, if these are also capitalized or are prepositions (see table 1).

Hypotheses Processing

In this section of the procedure, hypotheses are generated and evaluated. A hypothesis may consist of two adjacent capitalized words or a preposition with a capitalized word. These hypotheses are evaluated on the basis of all occurences of the second word found in the corpus.

A hypothesis of two capitalized words is rejected, if

1. the left word is already in the PN-lexicon

2. the right word is an inflected form which is not possible with PNs.

All other hypotheses are analyzed in the following way. If the left word is a MCpot or a derived form

of a MCpot, then the right word is a proper name. For example "Senatspräsident Spadolini" is analyzed as proper name "Spadolini" with the apposition "Senatspräsident" which is derived from the MCpot "Präsident". The hypothesis is also accepted if the right word has a genitive ending and occurs without this ending in the corpus, because only proper names may occur in such constructions, as in (15). Normal nouns have to be accompanied by an article, as in (16).

(15) die Strategie *Frankreichs*
     (the strategy of *France*)

(16) die Strategie des Mörders
     (the strategy of the murderer)

A hypothesis of a preposition and a capitalized word is rejected, if the capitalized word

1. is a potential minimal context

2. is followed by a genitive article

3. is followed by a past participle.

The latter two conditions exclude such constructions ("feste Syntagmen"), as in:

(17) aus Anlaß des
     (on the occasion of)

(18) in Kauf genommen
     (accepted)

In addition, it is checked whether we have a construction like "zu Olims Zeiten", i.e., whether the capitalized word has a genitive ending and is followed by a capitalized word. For example, we found the following proper names:

(19) in *Lafontaines* Worten
     (in the words of *Lafontaine*)

(20) in *Stoltenbergs* Bilanz
     (in *Stoltenberg's* the balance sheet)

(21) gegen *Hitlers* Ermächtigungsgesetz
     (against *Hitler's* Enabling Act)

All resulting hypotheses are evaluated by another procedure which takes into account the AVL–tree containing all capitalized words together with the distributional information described above. Because the corpus is very small and often there is only one occurence of a word, this information is not very reliable and therefore error–prone. This could be improved by the application of the procedure to a very large corpus (several million words). At this point, it is only checked whether the right word occurs with an article (a clue for a normal noun) and whether it often occurs with other capitalized words or prepositions (a clue for a proper name). Proper names are normally not used with articles with the exception of ones – mostly cases place names and institutional names – which always occur with an article (e.g. "die Türkei", "die Vereinigten Staaten"). So, this method has to be used carefully.

The processing of hypotheses is iterated until no more proper names can be found (pn_new = 0), because new proper names supply new contexts and new contexts may supply new proper names.

Tagging

In order to tag the proper names collected in the EN–lexicon, it is necessary to run through the corpus for a last time. All words listed in the EN–lexicon are tagged as proper names.

The procedure of proper name tagging was implemented in C under UNIX.

4 Evaluation

The first half of the corpus was used to develop the procedure, the second half served for an evaluation. For the evaluation, all proper names in the second corpus half were manually tagged and (manually) compared to the result of the automatic tagging procedure applied to this corpus part, i.e., to a corpus of 25,000 words. Of the 1300 proper name tokens 461 occurrences were not recognized, 30 text words were wrongly tagged as proper names. This corresponds to a recognition rate of about 65% (counting errors not excluded). In order to provide background for this figure, some of the problems are discussed here in more detail.

The preprocessing module could be improved by enlarging the MC–lexicon with a list of most frequently used first names, for example. For the recognition of non–German proper names, it could be possible to add non–German titles and forms of address as well. The latter were also found in the corpus (e.g. *Captain Alan Stephenson, Lord Carrington*).

At the Moment, first names are collected in the MCpot–lexicon if they are used attributively to a surname already recognized. This is in contrast to the approaches of (Fleischer, 1989) and others ((Wimmer, 1973), (Kalverkämper, 1978)), who analyze first names and surnames as a unit. One reason for this is that only the surname can be inflected, as in (22). But as this also applies to titles, as in (23), the reason does not hold.

(22) *Peter Müllers* Auto
   (the car of *Peter Müller*)

(23) Minister *Wörners* Rede
   (the speech of minister *Wörner*)

A better argument is that constructions of first name and surname cannot be expanded, e.g., as loose appositional constructions.

The procedure of proper name tagging described here is not able to recognize multi–word proper names because only two adjacent capitalized words (apposition + proper name) are examined. Table 2 shows an excerpt of unresolved hypotheses in which some multi–word proper names consisting of first name and surname (*Albrecht Müller, Angelika Beer, Harry Ristock, Ruth Winkler, Josef Felder, Gabi Witt, Florian Gerster, Sepp Binder, Kurt Schumacher*), of normal nouns (*(das) Deutsche Rote Kreuz, Kleine Brogel, Ewige Lampe*) and of some non-German proper names (*Alan Stephenson, (Canadian) Air Group, Central Enterprise, Frecce Tricolori, Standardisation Agreement, Acrobatic Full Scale*) are found.

The non-German proper names are often put in quotation marks, so this could be an additional criterion for the hypotheses evaluation, but cases in which quotation marks are used to emphasize or to cite one or more words must be excluded (24).

(24) die FDP warnt vor "Panikmache"
   (the FDP warns of "panic mongering")

Multi–word proper names consisting of normal nouns or mixed of normal nouns, adjectives, articles, prepositions and proper names constitute a major problem. Apart from the fact that adjectives and prepositions belonging to a proper name are capitalized, some of these proper names (25) behave like normal nouns, i.e., they are inflectional and take an article, but some do not (26)-(28). The latter are mostly used with an introductory apposition and often put in quotation marks. For one it is difficult to determine which constituents belong to the proper name, and which do not when the construction can be modified and reduced as well (e.g. *Vereinigte Staaten von Amerika, die Staaten, die Bundesrepublik, Deutschland*). Under the more distributional analysis described here, it is not possible to recognize them and no easy solution is possible. In secondly place, it is possible to recognize them if we know the minimal context (here *Luftwaffenbasis, Gasthaus, Straße*), which may be resolved if we use a very large corpus, and if we consider more than one following word and existing quotation marks.

Text	Hypothesis
1	Militaerflughafen Rhein-Main
2	Dutzend Personenwagen
2	Captain Alan
2	Alan Stephenson
6	Mitte April
7	Metern Abstand
11	Fraktionskollege Albrecht
11	Albrecht Mueller
12	Kanadische Luftwaffendivision
12	Air Group
12	Hochleistungsflugzeug F-18
13	Central Enterprise
13	Central Enterprise
14	Central Enterprise
22	Frecce Tricolori
22	Deutsche Rote
22	Rote Kreuz
22	Dutzend Demonstranten
22	Autobahnzufahrt Frankfurt-Sued
22	Luftsportgruppe Breitscheid/Haiger
23	Kleine Brogel
24	Fraktionskollegin Angelika
24	Angelika Beer
25	Ende September
27	IG Metall
27	Harry Ristock
27	Lehrerin Ruth
27	Ruth Winkler
28	Regierung Kohl
28	Prozent Kandidatinnen
30	Leitende Oberstaatsanwalt
30	Oberstaatsanwalt Sattler
32	Frecce Tricolori
34	Geburtstag Bert
34	Josef Felder
34	Gabi Witt
34	Ewige Lampe
34	Museumsdorf Muehlendorf
34	Florian Gerster
34	Sepp Binder
34	Kurt Schumacher
35	Standardisation Agreement
35	Standardisation Agreement
35	Acrobatic Full
35	Full Scale
36	Frecce Tricolori
36	Frecce Tricolori
36	Frecce Tricolori
36	Demokratische Proletarier
37	IG Metall
37	IG Chemie
37	IG Bergbau
39	Kanzleramt Erwaegungen
56	Partei Ernst
96	Bundespartei Stellung

Table 2: unresolved hypotheses (excerpt)

(25) *die Vereinigten Staaten* und *die Bundesrepublik Deutschland*
(*the United States* and *the Federal Republic of Germany*)

(26) auf der nordbelgischen Luftwaffenbasis *Kleine Brogel*
(at the North Belgian air force base *Kleine Brogel*)

(27) ein Teil von ihnen geht [...] ins Gasthaus "*Ewige Lampe*"
(some of them go to the inn "*Ewige Lampe*")

(28) ich habe in der Straße "*Am Mariahof*" gewohnt
(I have lived in the street "*Am Mariahof*")

Some of the remaining hypotheses in Table 2 are noun pairs consisting of quantity terms and normal nouns (29)-(31) or constructions with month names (32). Quantity terms could be excluded by an exception list and month names could be added to the EN–lexicon from the start.

(29) ein Dutzend Personenwagen/Demonstranten
(a dozen automobiles/demonstrators)

(30) mindestens vierzig Prozent Kandidatinnen
(at least 40 per cent candidates)

(31) nach Metern Abstand
(after a distance of some metres)

(32) Mitte April/Ende September
(in the middle of April/at the end of September)

But some of the remaining hypotheses are the result of a free German word order, often observed in sentences with support verb constructions (34: *Ernst machen mit* (to be serious about), 35: *Stellung beziehen gegen* (to take a stand against)). The hypotheses 'Kanzleramt Erwägungen' in sentence (33) could be ruled out if the form 'Erwägungen' was analyzed as a non–possible inflection form of a proper noun and therefore as a normal noun. This was not performed by the morphological analysis[1], because there were no occurrences of 'Erwägung' without a plural ending in the corpus. This could be improved by the use of a very large corpus or a powerful morphological analyzer (e.g. GERTWOL, (Koskenniemi and Haapalainen, 1994)). The support

---

[1]The analysis is based on a very simple mechanism: inflectional endings which are not possible for proper names are removed from the word under consideration, and the remaining form is searched for in the corpus. If successful, the word cannot be a proper name and the hypothesis is rejected; if not, the hypothesis is kept.

verb constructions could be excluded if we look for typical verbs used in such constructions (*machen, bringen, nehmen, ...*).

(33) ... war bekanntgeworden, daß im Kanzleramt Erwägungen [...] stattfänden, wie ...
(... became known that the chancellorship takes into consideration ...)

(34) ... wenn seine Partei Ernst macht mit ...
(... if his party gets serious about ...)

(35) ... indem man [...] gegen die Bundespartei Stellung bezieht
(... while taking a stand against the federal party)

Most of the incorrectly tagged proper names are the result of the hypotheses processing, because the corpus is too small. For example, the evaluation of the hypothesis 'ohne Rücksicht' (with no consideration) provides 'Rücksicht' as proper name, because it also occurs with the preposition 'aus' (from), which is frequently used with place names and never occurs with an article, but its frequency is only 4. This is not representative for a reliable conclusion and it is hoped that a very large corpus would allow for a better analysis.

5 Conclusions and Future Perspectives

Most of the known statistically based tagging systems are confronted with the problem of proper name tagging. In German the problem is not only restricted to the disambiguation of sentence–initial words but also occurs with sentence–internal capitalized words. The procedure of proper name tagging described here makes use of a database of definite minimal contexts as a starting point for an analysis which takes into account both morphological and syntactic properties of proper names. Furthermore, this local analysis is supported by a global analysis regarding all occurrences of capitalized words in the corpus. This global analysis should be improved by a larger corpus than the one used, and a more meaningful statistic procedure, like *mutual information* (Church and Hanks, 1990). However, the central idea of an incremental procedure for the collection of proper name contexts is encouraging. It is planned to include this proper name tagging in the German part-of-speech tagger LIKELY (Feldweg, 1993) developed in Tübingen to disambiguate all the remaining cases where the tagger could not decide between proper name or normal noun.

References

[Church and Hanks1990] K. Church and P. Hanks. 1990. Word association norms, mutual information and lexicography. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburgh*, pages 76–83.

[Church1988] Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143. Association for Computational Linguistics.

[Feldweg1993] Helmut Feldweg. 1993. Stochastische Wortartendisambiguierung für das Deutsche: Untersuchungen mit dem robusten System LIKELY. Technical report, Universität Tübingen, Seminar für Sprachwissenschaft.

[Fleischer1989] Wolfgang Fleischer. 1989. Struktur und Funktion mehrwortiger Eigennamen im Deutschen. In Friedhelm Debus and Wilfried Seibicke, editors, *Reader zur Namenkunde, Band I: Namentheorie*, pages 263–271. Georg Olms, Hildesheim; Zürich; New York. republished, Proceedings of the 13th International Congress of Onomastic Sciences 1981, 403–411.

[Garside et al.1987] R. Garside, G.L. Leech, and G. Sampson, editors. 1987. *The computational analysis of English*. Longman, London and New York.

[Hackel1986] Werner Hackel. 1986. Zum Verhältnis Eigenname und Apposition. *Namenkundliche Informationen*, (49):1–12.

[Kalverkämper1978] Hartwig Kalverkämper, editor. 1978. *Textlinguistik der Eigennamen*. Klett–Cotta, Stuttgart.

[Koskenniemi and Haapalainen1994] Kimmo Koskenniemi and Mariikka Haapalainen. 1994. Gertwol. *LDV-Forum*, 11(1):17–29.

[Koss1987] Gerhard Koss. 1987. Über das Lernen und Vergessen von Eigennamen. *Zeitschrift für Literaturwissenschaft und Linguistik*, 67:24–37.

[Koss1990] Gerhard Koss. 1990. *Namenforschung: Eine Einführung in die Onomastik*. Niemeyer, Tübingen. Germanistische Arbeitshefte, 34.

[Wimmer1973] Rainer Wimmer, editor. 1973. *Der Eigenname im Deutschen: Ein Beitrag zu seiner linguistischen Beschreibung*. Niemeyer, Tübingen. Linguistische Arbeiten, 11.

[Wirth1983] Nikolaus Wirth, 1983. *Algorithmen und Datenstrukturen*, page 244 ff. Teubner, Stuttgart, third edition.